Audio-Visual Affect Recognition

Zhihong Zeng, Jilin Tu, Ming Liu, Thomas S. Huang, Brian Pianfetti, Dan Roth, and Stephen Levinson

***Abstract*—The ability of a computer to detect and appropriately respond to changes in a user's affective state has significant implications to Human–Computer Interaction (HCI). In this paper, we present our efforts toward audio–visual affect recognition on 11 affective states customized for HCI application (four cognitive/motivational and seven basic affective states) of 20 nonactor subjects. A smoothing method is proposed to reduce the detrimental influence of speech on facial expression recognition. The feature selection analysis shows that subjects are prone to use brow movement in face, pitch and energy in prosody to express their affects while speaking. For person-dependent recognition, we apply the voting method to combine the frame-based classification results from both audio and visual channels. The result shows 7.5% improvement over the best unimodal performance. For person-independent test, we apply multistream HMM to combine the information from multiple component streams. This test shows 6.1% improvement over the best component performance.**

***Index Terms*—Affect recognition, affective computing, emotion recognition, multimodal human–computer interaction.**

I. Introduction

Until recently, access to changes in affect which are crucial in human decision making, perception, interaction, and intelligence were inaccessible to computing systems. Emerging technological advances are enabling and inspiring the research field of "affective computing" which aims at allowing computers to express and recognize affect [11].

The work in this paper is motivated by the ITR project [14]. The goal of this project is to contribute to the development of multimodal human-computer intelligent interaction environment. An educational learning environment was used as a test-bed to evaluate the ideas and tools resulting from this research. This test-bed focused on using Lego gears to teach math and science concepts to upper elementary and middle school children. The project focuses on using proactive computing to achieve two ends. The first is to help children explore and understand a variety of phenomena ranging from mathematic ratios to advanced concepts of mechanical advantage and least common multiples. The second goal of the project is to support and prolong a student's interest in the activities while also promoting a high level of student engagement. This is accomplished through a multimodal computer learning environment that uses audio–visual sensors to recognize the student's affective states and to proactively apply appropriate context specific tutoring strategies.

Multimodal sensory information fusion is a process that enables human ability to assess emotional states robustly and flexibly. The psychological study [12] indicated that people mainly rely on facial expressions and vocal intonations to judge someone's affective states. To more accurately simulate the human ability to assess affect, an automatic affect recognition system should also make use of multimodal data.

Automatic emotion recognition has been attracting researchers from a variety of different disciplines [1], [7]. However, most current automatic emotion recognition approaches are unimodal: information processed by the computer system is limited to either face images [8], [9], [17] or speech signals [19], [23]–[26]. Relatively little work has been done in researching multimodal affect analysis [2]–[5]. In addition, most current emotion recognition approaches focus on the primary or basic emotions which are primitive emotions and could only involve the low-level brain processing.

Recently, there has been an increasing number of studies to explore multimodal emotion recognition [20]–[22]. One notable work was done by Fragopanagos and Taylor in 2005 [22] who explored multimodal emotion recognition on natural emotion data. They use the Feeltrace tool [18] to label the data and use a neural network architecture to handle the fusion of different modalities. In their work, due to considerable variation across four raters, it is difficult to reach a similar assessment with the FeelTrace labels. In addition, their recognition of four emotion classes which are quadrant matching in activation-evaluation space leads to some loss of information by collapsing the structured, high dimensional space of the possible emotional states to a rudimentary four-class space.

In human–computer interaction for education, we are more concerned some cognitive and motivational states which are more subtle and sophisticated. These states could require high-level brain processing, and even conscious awareness. Recognizing these states (e.g., interest, boredom, frustration, and puzzlement), the computer is able to proactively apply appropriate context specific tutoring strategies (e.g., encouragement, transition/guidance, and confirmation).

In this paper, we expand the number of affective states researched beyond seven basic emotions to include four cognitive/motivational states (puzzlement, interest, boredom, and frustration). These cognitive/motivational states can give us insight into the progress, strategies, and engagement associated during the course of learning.

Due to the fact that it is quite difficult to obtain sufficient natural audio-visual affect material for this fine-grained emotion recognition, and there is no objective labeling system to annotate this audio-visual data, we test our algorithm on required affect data from 20 nonactor people. Although the subjects displayed affect expression on request, no instruction was given as to how to perform the emotional states. The performed emotions are based on the subject's individual perception of prototypical emotional responses. Thus, the analysis of this data can provides some insight what features people are likely to use and how they use them to display their emotions.

II. Database

A large-scale database was collected [15] that is customized for multimodal affect recognition research for human-computer interaction applications. Eleven affect categories were used which include seven basic affects (i.e., happiness, sadness, fear, surprise, anger, disgust, and neutral), and four cognitive/motivational affects (i.e., interest, boredom, puzzlement, and frustration).

The 20 subjects (ten females and ten males) in our database consist of graduate and undergraduate students, faculty, and staff in a higher-education institution. These subjects were from a variety of backgrounds and fields of interest. The subjects displayed affect expression without instruction, based on the subjects' individual perception of prototypical emotional responses. Each subject was required to pose each facial expression while speaking the specific appropriate sentence three times.

Manuscript received July 6, 2005; revised June 24, 2006. This work was supported by Beckman Postdoctoral Fellowship and National Science Foundation: Information Technology Research Grant 0085980. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jie Yang.

The authors are with Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: zhzeng@ifp.uiuc.edu; jilintu@ifp.uiuc.edu; mingliu1@ifp.uiuc.edu; huang@ifp.uiuc.edu; bpianfet@uiuc.edu; danr@cs.uiuc.edu; sel@ifp.uiuc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2006.886310

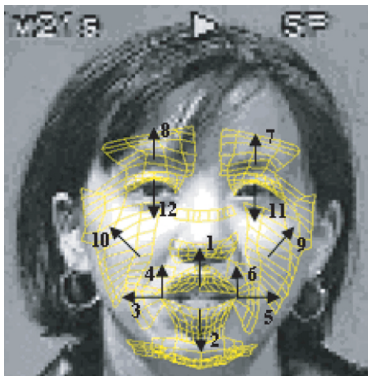

Fig. 1. The 12 facial motion units.

Every emotion expression sequence lasts from 2 to 6 s. The average length of expression sequences is 4 s.

Subjects appeared to express affects more naturally while reciting appropriate sentences than without speech. Specifically, subtle differences between the facial expressions of the four cognitive/motivational states were difficult to display without accompanying speech. However, on the other hand, speaking reduces the discriminability among facial expressions at different affective states. That is further analyzed in Section III.

During our labeling, start and end points of each affect expression were determined by speech energy which is easy to detect.

III. Feature Extraction

In this section, we present the techniques used for extracting the facial features and prosodic features.

A. Facial Affective Feature Extraction

A tracking algorithm called Piecewise Bezier Volume Deformation (PBVD) tracking [16] is applied to extract facial features in our experiment.

This face tracker uses a 3-D facial mesh model which is embedded in multiple Bezier volumes. The shape of the mesh can be changed with the movement of the control points in the Bezier volumes. That guarantees the surface patches to be continuous and smooth. In the first video frame (frontal view of a neutral facial expression), the 3-D facial mesh model can be constructed by manual or automatic selection [10] of landmark facial feature points. Once the model was fitted, the tracker can track head motion and local deformations of the facial features. At the current stage, only local deformations of facial features are used for affect recognition. These deformations are measured in terms of magnitudes of 12 predefined motions of facial features, called Motion Units (MUs), which are shown in Fig. 1. The outputs of the face tracker corresponding to 12 MUs and their derivatives are used as facial feature candidates for later feature selection analysis. The face tracker outputs 30 frames/s.

In our dataset, we notice two aspects of facial expression with speaking: on one hand, speech signals provide the affect information which complements facial expression; on the other hand, speaking influences facial expression so that the discriminability among facial expressions at different affective states decreases. The movements of facial features are related to both affective states and content of speech. Specifically, the mouth features corresponding to MU1-MU6 are influenced more by speech content than by affective states. That causes a decrease in the performance of face-only affect recognition.

In this paper, we apply a smoothing method to reduce the influence of speech on facial expression based on the assumption that the influence of affects on facial expressions is relatively more persistent than that of speech. The smoothing features are calculated by averaging facial features at consecutive frames to reduce the influence of speech on facial expression. In our experiment, we use ten frames as the length of the smoothing window.

B. Prosodic Feature Extraction

In our work, we use prosodic features which are related with the way the sentences are spoken. For prosodic feature extraction, a signal processing systems named formant, developed by Entropic Research Laboratory, is used. For each frame of sampled data, we choose 20 prosodic features, including pitch F0, RMS energy, formants F1-F4 and their bandwidths, and all of their corresponding derivatives as our prosodic feature candidate.

IV. Person-Dependent Affect Recognition

The personal-dependent recognition is first evaluated on these 20 subjects in which the training sequences and test sequence were taken from the same subject. For each subject in this test, one sequence among three sequences of each affective state is used as the test sequence, and the remaining two sequences of this state are used as training sequences. Accordingly, for the classification of 11 affective states, there are 11 sequences for test and 22 sequences for training. This test is repeated three times, each time leaving a different sequence of each state out.

We use a classifier named Sparse Network of Winnow (SNoW) [6] to measure the recognition accuracy of each visual or audio frame. Then a voting method is done in each expression sequence to obtain audio-visual fusion results.

A. Feature Selection

Out of these 24 facial features and 20 prosodic features, we want to identify those that contribute more in the classification. Because it is forbiddingly time-consuming to exhaustively search the subset of features that give best classification, we apply backward elimination and forward selection to identify the subset that are more important in distinguishing affects. Due to different sampling rate and asynchrony of audio and visual modalities, we do feature selection on individual modality.

1) Facial Feature Selection: The facial feature selection results are shown in Figs. 2 and 3. Fig. 2 display the curve of average accuracy versus the number of features in backward selection (top) and forward selection (bottom). Both of backward and forward feature selection results show that with the increase of the feature number n, accuracies roughly increase when $n < 12$, then keep almost 50% when $12 < n < 17$, finally gradually decrease when $n > 17$. Thus, it can be good to choose about 12 as the feature number.

Fig. 3 shows the frequencies (the number of occurrences) of facial features in backward (top) and forward (bottom) feature subsets. The x axis represents the feature number in which 1–24 represent MU1, MU2, MU3, MU4, MU5, MU6, MU7, MU8, MU9, MU10, MU11, MU12, Δ MU1, Δ MU2, Δ MU3, Δ MU4, Δ MU5, Δ MU6, Δ MU7, Δ MU8, Δ MU9, Δ MU10, Δ MU11, and Δ MU12, respectively. Features with high frequency are considered to contribute more to affect recognition. The results suggest that the important features for our classification are features 1-12. Among them, features 7 and 8, which are vertical movement of right brow and left brow individually, show the most discriminant ability. Comparatively, the derivatives of 12 motion units contribute less to the recognition.

In addition, the results suggest that although speech influences on mouth movement, the features around mouth still have some contribution on affect recognition.

Based on the results of facial feature selection analysis, we choose 12 facial motion units as the following classification features based

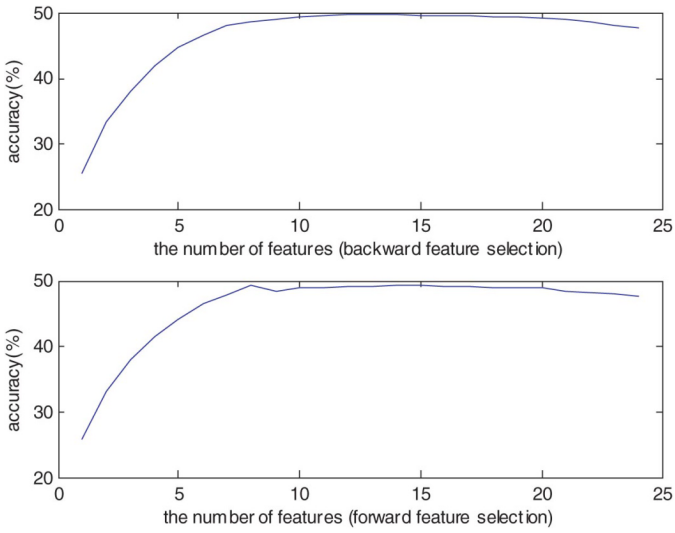

Fig. 2. Classification accuracy versus the number of facial features. Top is backward selection and bottom is forward selection.

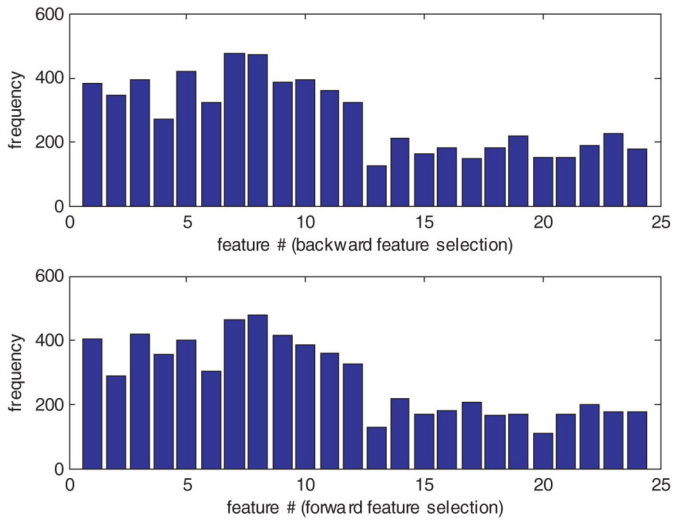

Fig. 3. Frequencies (the number of occurrences) of facial features in backward and forward feature subsets. Top is backward selection and bottom is forward selection.

on which we get 49.64% frame-based recognition rate for 11 affective states. Then, the facial smoothing method is applied to get 57.25% accuracy which has 7.61% improvement over the nonsmoothing classification.

2) Prosodic Feature Selection: The prosodic feature selection results are shown in Figs. 4 and 5. Fig. 4 displays the curve of average accuracy versus the number of audio features in backward selection (top) and forward selection (bottom). Both backward and forward feature selection results show that with the increase of audio feature number n, the accuracies increase when $n < 12$, then keep about 45% when $12 < n < 17$, finally decrease when $n > 17$. Thus, it could be good to choose about 12 as the number of features.

Fig. 5 shows that the frequencies (the number of occurrences) of audio features in backward (top) and forward (bottom) feature subsets. The x axis represents the feature number in which 1–20 represents F0, E, Δ F0, Δ E, F1, B1, F2, B2, F3, B3, F4, B4, Δ F1, Δ B1, Δ F2, Δ B2, Δ F3, Δ B3, Δ F4 and Δ B4, respectively. The more discriminant features include F0, E, Δ F0, Δ E, F1-F4, and B1-B3. Among them, features 1 and 2, which are pitch and energy individually, are the two most

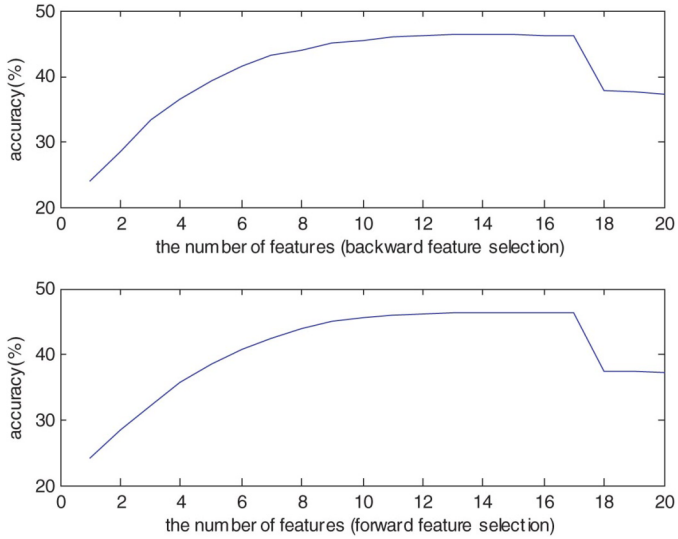

Fig. 4. Classification accuracy versus the number of audio features. Top is backward selection and bottom is forward selection.

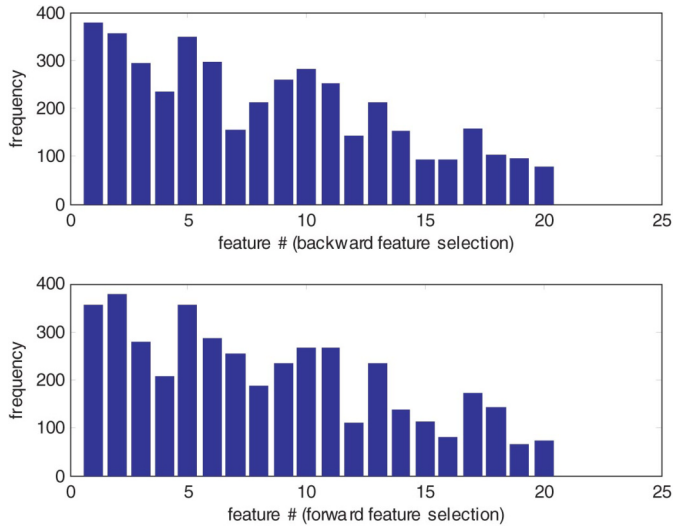

Fig. 5. Frequencies (the number of occurrences) of audio features in backward and forward feature subsets. Top is backward selection and bottom is forward selection.

important features. That agrees with the previous report [19]. Comparatively, the derivatives of formants and derivatives of bandwidths contribute less to the recognition.

Based on the results of audio feature selection analysis, we choose the 11 most discriminant audio features as the following recognition features based on which we get 44.66% frame-based recognition rate for 11 affective states.

B. Frame-Based Classification

In this paper, we apply the SNoW classifier as our frame-based classifier which was developed by Roth [6]. In the input layer of SNoW, raw measurements are transformed to a higher dimensional space of binary features. Consequently, the connections from transformed feature nodes to each of the output target nodes will be sparse, and the classes which are not linearly separable in the original space of raw measurements could become more linearly separable in the transformed higher dimensional space. In this work, we divide each feature space into ten bins through histogram method in which samples are put into ten disjoint sets of almost same sample size, activating a certain binary feature if a raw affective feature falls into the associated bin. In the

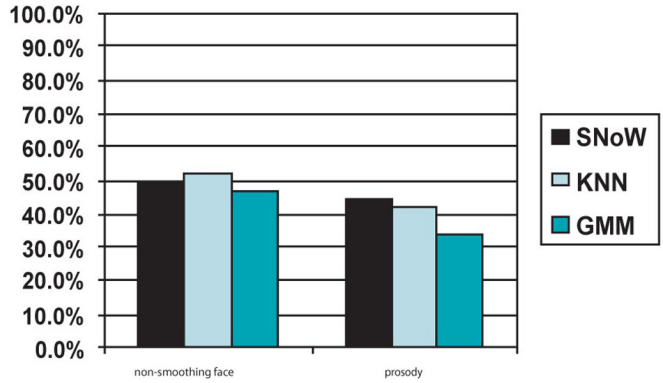

Fig. 6. Recognition comparison of SNoW, KNN, and GMM. The y axis is the recognition accuracy. The left parts are results in nonsmoothing face modality and right parts are results in prosody modality.

output layer, the output of each of target nodes representing classes is the weighted sum of the transformed features. In the learning procedure, several rules can be used to update connection weights, including winnow, perceptron, and Naïve Bayes [13]. In this paper, we choose SNoW-Naïve-Bayes (SNoW-NB).

In our experiments, we compared SNoW-NB with two classical classifiers, k nearest neighbor (KNN) and Gaussian mixture model (GMM). For KNN, we tried $k = 5$, 9, 13, and 17, and k is selected as 9 because of its best performance. For GMM, we select the number of component as 2, and use Expectation–Maximum (EM) to estimate the Gaussian mixture parameters. The recognition comparison of SNoW, KNN, and GMM in individual audio and visual modalities is shown in Fig. 6.

In Fig. 6, left parts are recognition results of nonsmoothing face modality, and right parts are results of prosody modality. It shows that in our application, GMM with two components has the poorest performance, and SNoW and KNN have the comparative performance. Unlike GMM, SNoW, and KNN do not assume any model, and have few parameters to estimate. SNoW is mainly related to the number of bins and KNN is mainly related to the number of nearest neighbors (k). But KNN is computationally expensive and needs to keep all the instances in the memory. Thus, we choose to use SNoW in our work.

C. Bimodal Fusion

In this section, we explore to combine these two modalities to achieve better recognition performance. In our work, the face modality outputs sets of facial features at a rate of 30 Hz. The prosodic modality outputs sets of prosodic features at a rate of 90 Hz. We choose decision-level fusion which combines multiple unimodal classifier outputs for final affective state classification. We apply a voting method in every expression sequence for bimodal fusion. In detail, every classification result from the face-only and prosody-only classifiers is treated as one vote. And, in every expression sequence, the system computes the vote number of every affective state. The affective state with the largest number of votes is designated as the final classification result.

Although the frame-level recognition accuracies of prosody-only and face-only modalities are not sufficient (only achieve about 50%) for our experiment, they are larger than random results (11 states have even priors, i.e., 9.09%). Consequently, the chance that the correct label wins in one expression sequence is greatly increased. In addition, two modalities together provide more votes than each of single modalities. Thus, bimodal voting method results to the improvement of the final affect recognition performance. The average affect recognition results are shown in Fig. 7. In Fig. 7, from left to right, the bars represents the frame-based accuracies of nonsmooth face modality and smooth face modality, sequence-based accuracy of face voting, frame-based accuracy of prosody modality, sequence-based accuracies of prosody voting, and bimodal voting. That shows that the average sequence-based accuracy of our bimodal fusion is 96.30%, 7.5% more accurate than the best of unimodal sequence-based performance.

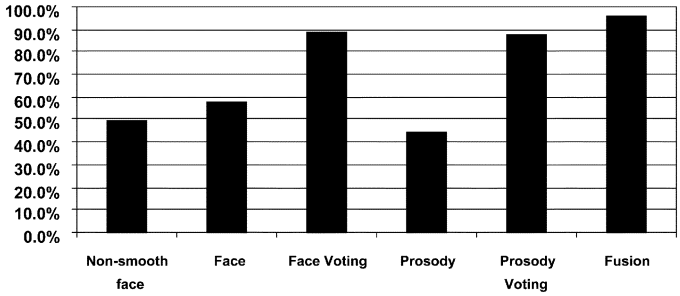

Fig. 7. Affect recognition results. From left to right, the bars represents the frame-based accuracies of nonsmooth face modality and smooth face modality, sequence-based accuracy of face voting, frame-based accuracy of prosody modality, sequence-based accuracies of prosody voting, and bimodal voting.

V. Person-Independent Affect Recognition

In this section, we explore person-independent affect recognition in which testing data and training data are from different subjects. Thus, the variation of the data is more significant and classification is more challenging than person-dependent recognition. For this test, we apply leave-one-subject-out cross-validation. In this way, all of the sequences of one subject are used as the test sequences, and the sequences of the remaining 19 subjects are used as training sequences. The test is repeated 20 times, each time leaving a different person out. For every affective state, there are $3^*20 = 60$ expression sequences. Therefore, there are totally $60^*11 = 660$ sequences for 11 affective states.

Regarding person-independent affect recognition, facial feature normalization is crucial because every subject has different physiognomy. To express an affect, different subjects will display different magnitudes of 12 MUs. To overcome this difference, the neutral expression for each person has been used as the normalization standard. In detail, for a given subject, the each of 12 MUs at every frame was normalized by the corresponding feature mean of the neutral expression of the same subject. After the feature vector of each frame is normalized, it is quantized into 19-size codebook by vector quantization (VQ).

The experimental results in the above feature selection showed pitch and energy are the most important factors in affect classification. Therefore, in the person-independent experiment, we only used these two audio features. Some prosody features, like frequency and duration of silence, could have implication in the HMM structure of energy and pitch.

The audio features also need normalization because they depend on the subject and recording condition. For a given subject, the pitch at every frame is normalized by the pitch mean of the neutral expression sequence of the same subject. The same is done for energy features to normalize amplitude change due to the speaker volume and the distance between a speaker and microphone. Then, the energy and pitch are quantized into 19-size codebook by vector quantization respectively.

We apply Multistream Hidden Markov Model (MHMM) for the person-independent recognition. In MHMM framework, the composite facial feature from video, energy and pitch features from audio are treated as three streams, and modeled by three component HMMs. We used the weighted summation to fuse the results from three component HMMs.

The performance of component HMMs from face-only, pitch-only, and energy-only streams, and MHMM fusion are shown in Table I. In MHMM, their corresponding weights of three component HMMs are set to be 2.5, 3.5, and 4, respectively. In Table I, face-only HMM gave the poorest performance (38.64%). The pitch-only HMM (57.27%) and energy-only HMM (66.36%) performed better than face-only HMM

TABLE I
PERSON -INDEPENDENT AFFECT RECOGNITION RESULTS

%	neut	happy	Sad	angry	disg	Surp	fear	frust	puzzle	inter	bore	**average**
face	45.00	40.00	31.67	58.33	43.33	43.33	51.67	18.22	41.67	25.00	26.67	**38.64**
pitch	91.67	41.67	55.00	63.33	68.33	41.67	61.67	35.00	71.67	61.67	38.33	**57.27**
energy	91.67	53.33	33.33	66.67	85.00	60.00	71.67	63.33	78.33	58.33	68.33	**66.36**
MHMM	96.67	53.33	66.67	76.67	90.00	56.67	71.67	58.33	81.67	81.67	63.33	**72.42**

but worse than MHMM (72.42%) because MHMM combine information of face, pitch, and energy which provide complementary information for recognition. This test shows 6.1% improvement over the best component performance.

For face-only HMM in Table I, the three cognitive states (frustration, interest, and boredom) has the lowest recognition rates. That shows that it is difficult to judge these subtle cognitive states if only using information of facial expression. Compared with facial features, the audio features (pitch and energy) provide more information in these states.

VI. CONCLUSION

An automatic affect recognizer has the potential to allow a computer to respond appropriately to the user's needs rather than simply responding to user's command. From this perspective, the computer can become an active participant in the HCI experience resulting in communications that are more natural, persuasive, and friendly. Proactive computing has the potential of having a significant application in helping children learn new skills and stay engaged longer with this type of computing system than traditional computer based learning approaches.

In this paper, we introduce our effort toward multimodal affect recognition on 11 affective states (four cognitive/motivational and seven basic affective states) of 20 nonactor subjects. To properly choose relevant features set, we apply feature selection approach to analyze the curve of average classification accuracy versus number of features, and feature ranks. A smoothing method is proposed to reduce the detrimental influence of speech on facial expression recognition. Our experimental results show that bimodal affect recognition can perform 7.5% in person-dependent test and 6.1% in person-independent test more accurately than best of the unimodal performance.

Multimodal recognition of human affective states is a largely unexplored and challenging problem. Our algorithm was only tested on a database of elicited emotional expressions which have the potential to differ from corresponding performances in natural settings. Consequently, additional experiments need to be performed on naturally occurring affect data. In order to do it, we have to face the difficulties to collect and label natural emotion data, and the technique limitation which is challenged by arbitrary head movement, hand occlusion, subtle and short lived facial expression, and nonoptimal imaging and audio condition.

ACKNOWLEDGMENT

The authors thank Dr. L. Chen for collecting the valuable data in this paper for audio-visual affect recognition.

REFERENCES

[1] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proc. IEEE*, vol. 91, pp. 1370–1390, Sep. 2003.
[2] L. Chen and T. S. Huang, "Emotional expressions in audiovisual human computer interaction," in *Proc. Int. Conf. Multimedia Expo*, 2000, pp. 423–426.
[3] L. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Int. Conf. Automatic Face Gesture Recognition*, 1998, pp. 396–401.
[4] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. Int. Conf. Automatic Face Gesture Recognition*, 2000, pp. 332–335.
[5] Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face ," in *Proc. ROMAN*, 2000, pp. 178–183.
[6] D. Roth, "Learning to resolve natural language ambiguities: A unified approach," in *Proc. Nat. Conf. Artificial Intelligence (AAAI)*, Jul. 1998.
[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, pp. 32–80, Jan. 2001.
[8] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, 2001.
[9] I. Essa and A. Pentland, "Facial expression recognition using a dynamic model and motion energy," in *Proc. Int. Conf. Computer Vision*, 1995, pp. 360–367.
[10] J. Tu, Z. Zhang, Z. Zeng, and T. S. Huang, "Face localization via hierarchical condensation with fisher boosting feature selection ," in *Proc. Computer Vision Pattern Recognition*, 2004.
[11] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.
[12] A. Mehrabian, "Communication without words," *Psychol.Today*, vol. 2, no. 4, pp. 53–56, 1968.
[13] A. J. Carlson, C. M. Cumby, N. D. Rizzolo, J. L. Rosen, and D. Roth, SNoW User Manual.
[14] [Online], Available: itr.beckman.uiuc.edu.
[15] L. S. Chen, Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction UIUC, 2000, Ph.D. dissertation.
[16] H. Tao and T. S. Huang, "Explanation-based facial motion tracking using a piecewise bezier volume deformation mode ," *Proc. Computer Vision and Pattern Recognition*, vol. 1.1, pp. 611–617, 1999.
[17] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vis. Image Understand.*, vol. 91, no. 1–2, pp. 160–187, Jul.-Aug. 2003.
[18] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'Feeltrace': An instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop on Speech and Emotion*, 2000, pp. 19–24.
[19] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," in *Proc. EUROSPEECH*, 2003.
[20] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 2005, vol. II, pp. 1085–1088.
[21] Y. Wang and L. Guan, "Recognizing human emotion from audiovisual information," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. II, pp. 1125–1128.
[22] F. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, pp. 389–405, 2005.
[23] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, pp. 5–23, 2003.
[24] D. J. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *Proc. 42nd Annu. Meeting of the Association for Computational Linguistics (ACL)*, Jul. 2004.
[25] R. Huber, A. Batliner, J. Buckow, E. Noth, V. Warnke, and H. Niemann, "Recognition of emotion in a realistic dialogue scenario," in *Proc. Int. Conf. on Spoken Language Processing*, 2000, vol. 1, pp. 665–668.
[26] T. Polzin, "Detecting Verbal and Non-Verbal Cues in the Communication of Emotion," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1999.